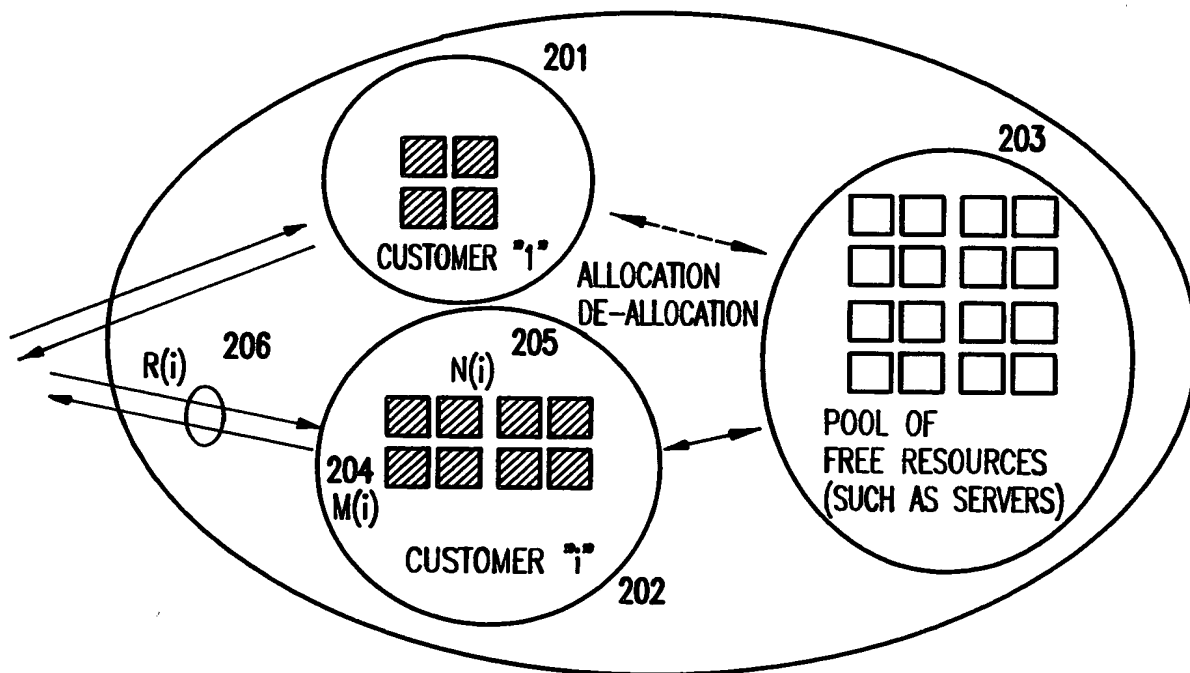


FIG.1



MEASURE $R(i)$, $N(i)$, AND $M(i)$
 COMPUTE $N_t(i)$ AND $R_t(i)$ FROM $M_t(i)$, $R(i)$, $N(i)$ AND $M(i)$,
 AND THEN IF NEEDED, MOVE $M(i)$ TO $M_t(i)$ BY EITHER CHANGING $N(i)$
 TO $N_t(i)$ OR CHANGING $R(i)$ TO $R_t(i)$.

FIG.2

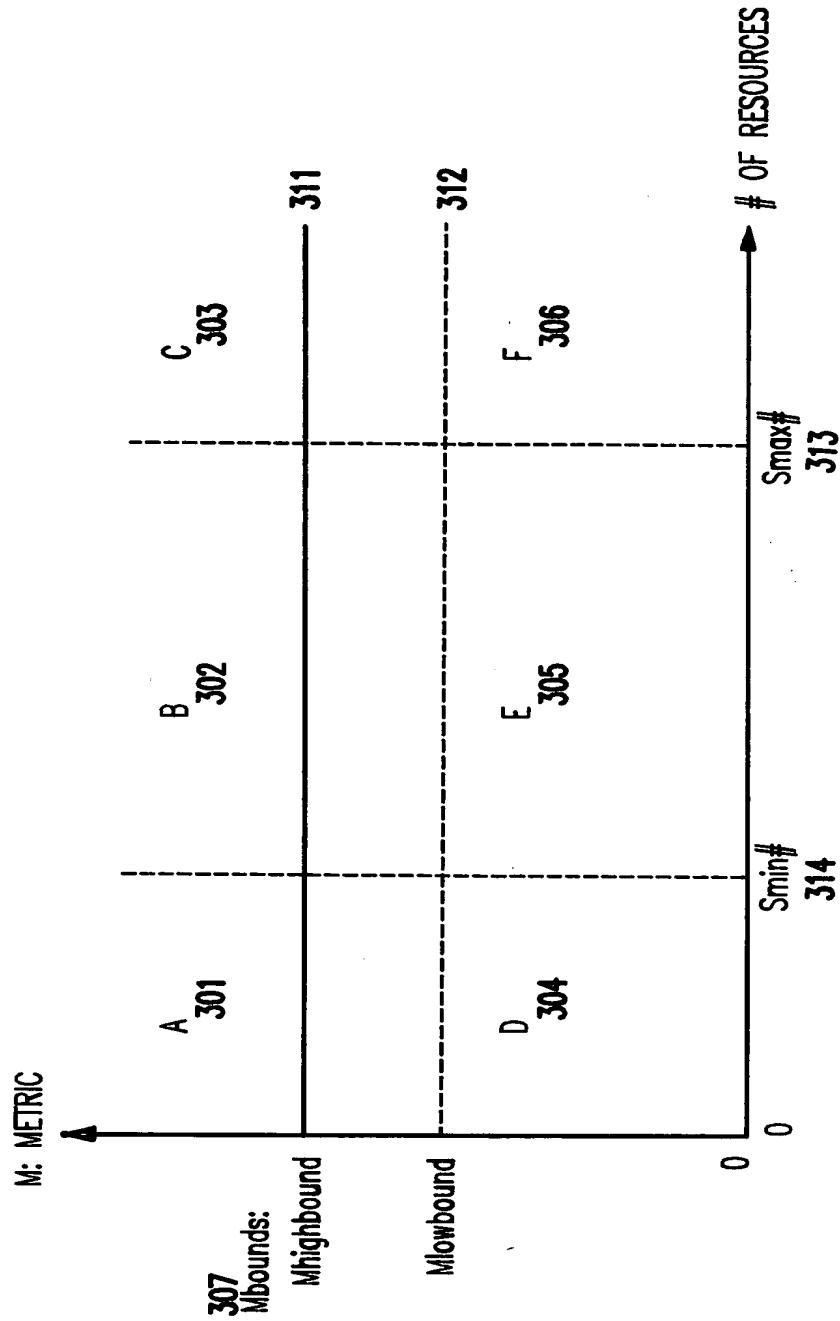
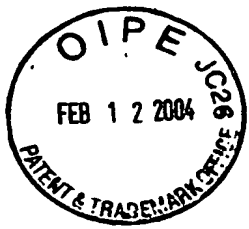


FIG.3

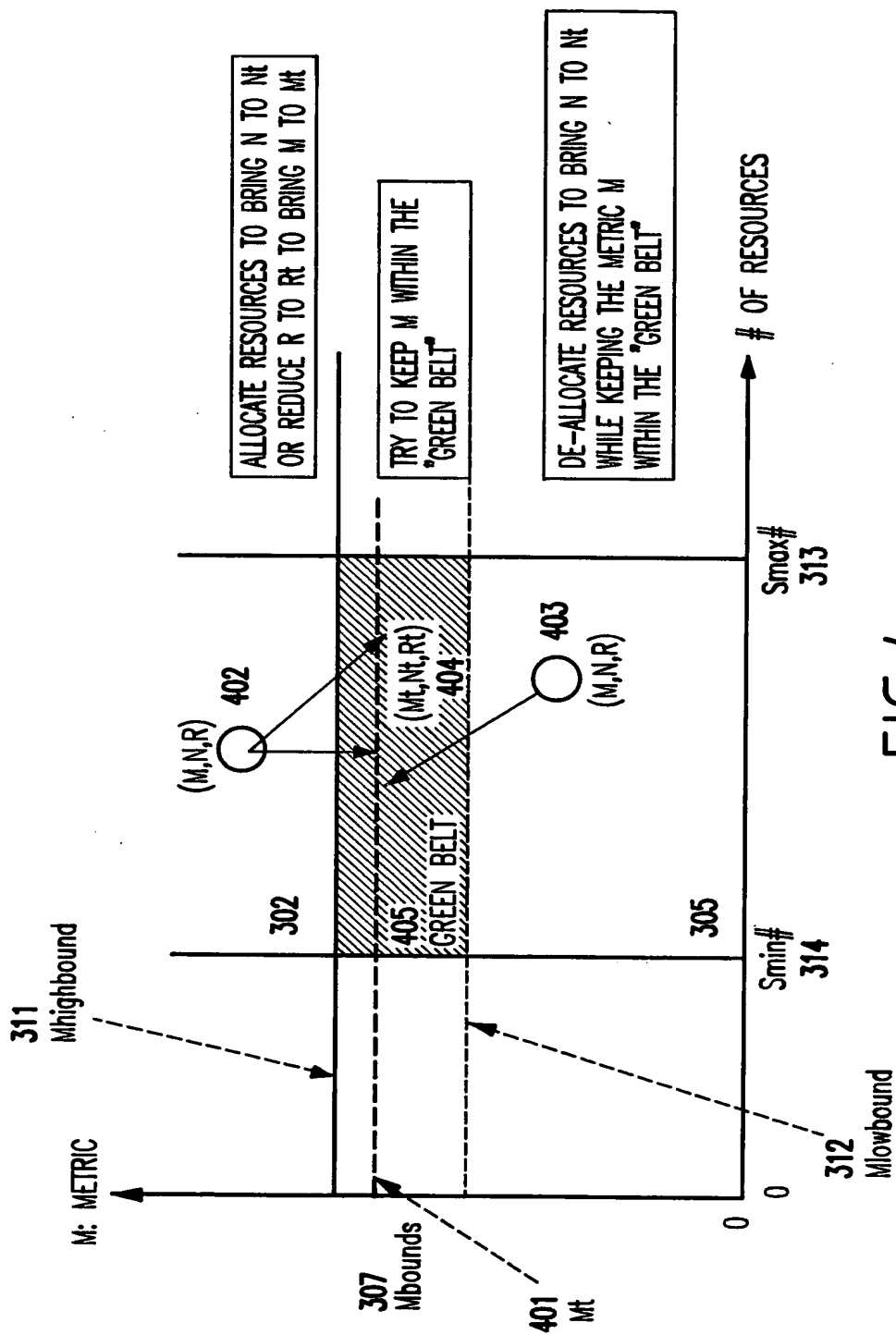


FIG. 4

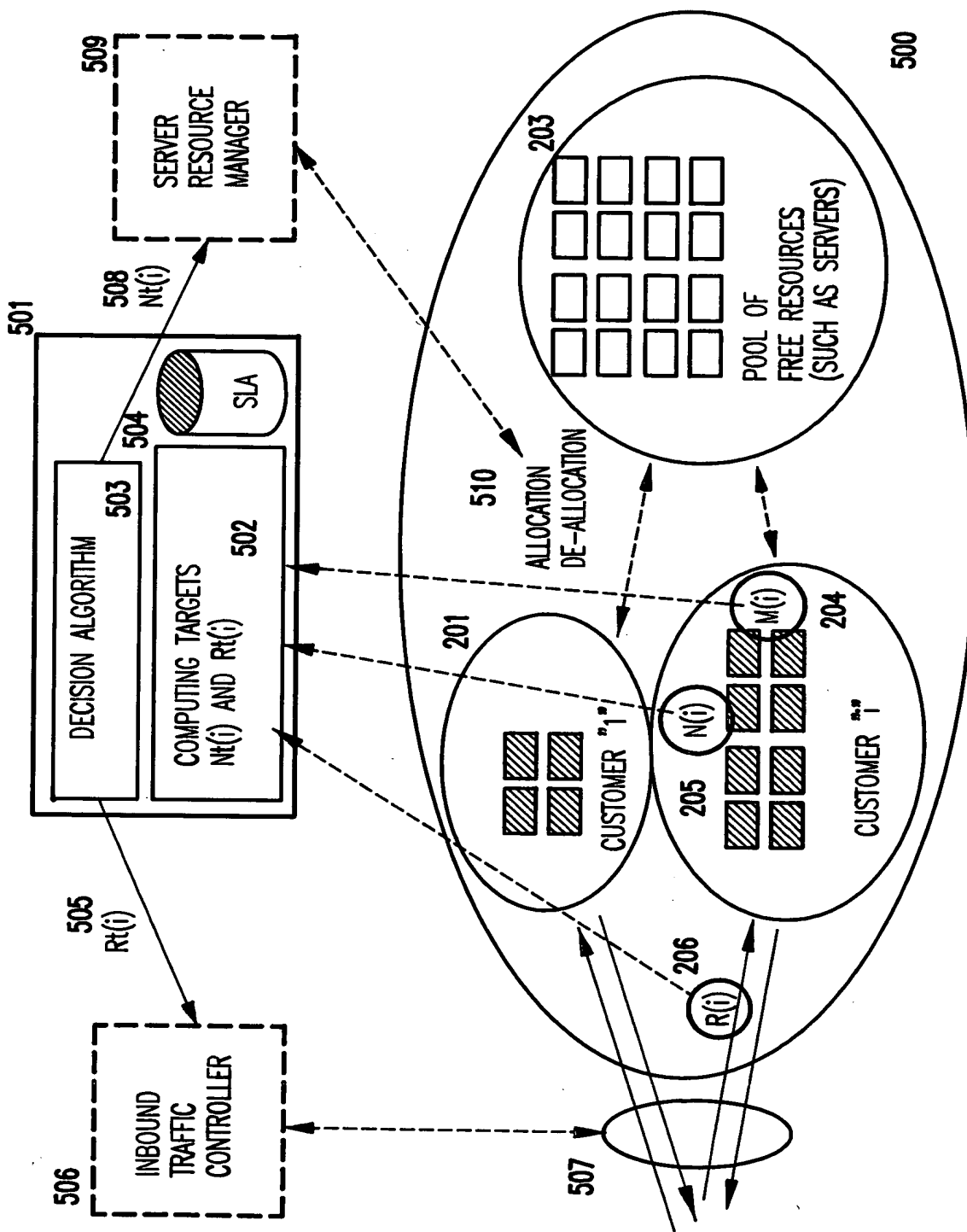


FIG. 5

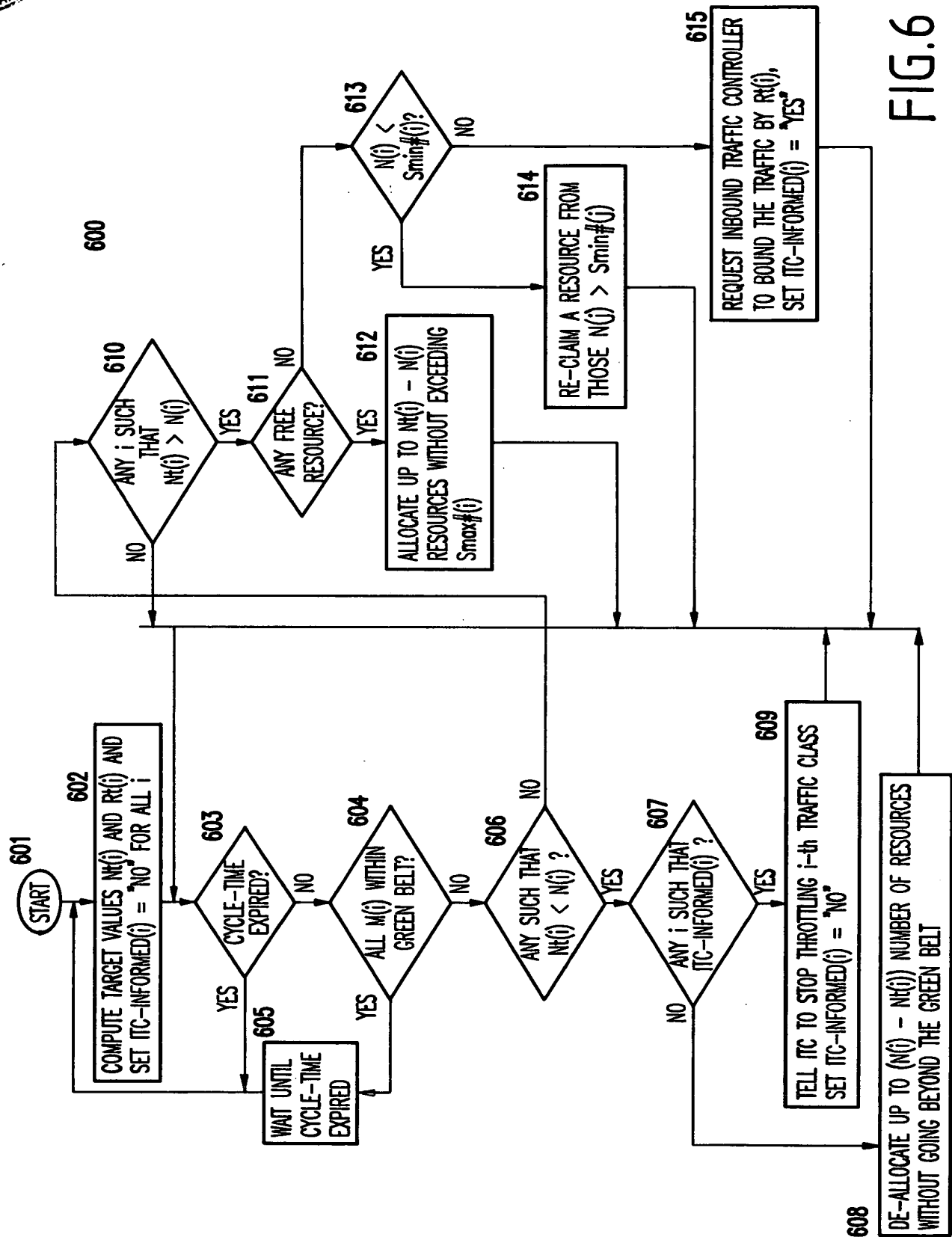


FIG. 6



Smin#(i)	: the amount of resources guaranteed for the i-th customer. This can be a vector.
Smax#(i)	: the maximum amount of service resources that could be made available to the i-th customer.
	This can be a vector.
Mbounds(i)	: the bounds on the service level metric.
	Each "bounds" consists of a pair, "highbound" and "lowbound."
Ubounds(i)	: the bound on the utilization of resources allocated to the i-th customer
Tbounds(i)	: the bound on the agreed upon average server response time for the i-th customer
T%bounds(i)	: the bound on the agreed upon server response time percentile for the i-th customer
(Smin#(i),Smax#(i),Mbound(i))	: the SLA supported by the invention
N(i)	: the number (or amount of) of resources currently allocated to the i-th customer.
R(i)	: the current inbound traffic rate for the i-th customer. This could be a vector when more than one type of traffic is defined for each customer.
M(i)	: the current value of the metric M for the i-th customer. This could be a vector.
	Examples are:
U(i)	: the current utilization of the allocated resources to the i-th customer
T(i)	: currently observed server response time averaging for the i-th customer
T%(i)	: currently observed server response time percentile for the i-th customer
Mt(i)	: the "target" (want to achieve) metric value for the i-th customer. Its dimension is same as the dimension of M(i).
	This is within the defined "green belt" which is the region within which M(i) is kept.
	Examples of Mt(i) are:
Ut(i)	: the target resource utilization when M = U,
Tt(i)	: the target average response time when M = T
Tt%(i)	: the target percentile response time when M = T%

Table 1

**For Utilization as Metric: $M = U$ and $Mt = Ut$**

The following relationships hold among various variables:

$$U(i) = C(i)R(i) / N(i), \text{ where } C(i) \text{ is a constant}$$

$$Ut(i) = C(i)R(i) / Nt(i), \text{ and}$$

$$Ut(i) = C(i)Rt(i) / N(i).$$

From the above and from the given values of $N(i)$, $R(i)$, $U(i)$, and the target value $Ut(i)$, $Nt(i)$ and $Rt(i)$ can be computed as follow:

$$Nt(i) = \text{CEILING} [N(i)U(i) / Ut(i)], \text{ and}$$

$$Rt(i) = \text{FLOOR} [R(i)Ut(i) / U(i)],$$

where CEILING gives the smallest integer exceeding and FLOOR gives the largest integer not exceeding.

Table 2

For Average Response Time as Metric: $M = T$ and $Mt = Tt$

S(i) :server "service" (or processing) time for the i-th customer, this can be computed from observing each individual server service time, or estimated from a queueing formula:

S(i) is a function of $\{T(i), R(i), N(i)\}$

If the cluster of servers is modeled by the M/M/m queueing system,

$$\mathbf{S(i)} = ((R(i)T(i)+N(i)+p\{N(i)\}) - \text{SQRT}((R(i)T(i)+N(i)+p\{N(i)\})^2 - 4R(i)T(i)R(i)) / 2R(i)$$

where $p\{m\}$ is the probability that there are m requests in the i-th customer's server cluster

For the M/M/m queueing model,

$$Tt(i) \sim S(i) + p\{Nt(i)\}S(i) / (Nt(i)-R(i)S(i))$$

$$Tt(i) \sim S(i) + p\{N(i)\}S(i) / (N(i)-Rt(i)S(i))$$

Therefore,

$$\mathbf{Nt(i)} = \text{CEILING} [R(i)S(i) + p\{Nt(i)\}S(i) / (Tt(i)-S(i))]$$

$$\mathbf{Rt(i)} = \text{FLOOR} [N(i)/S(i) - p\{N(i)\}/(Tt(i)-S(i))]$$

where $p\{m\}$ is the probability that there are m requests in the customer's server cluster

Table 3



For Percentile Response Time as Metric: $M=T\%$ and $Mt=Tt\%$

If $T\%(i) > T\%bound(i)$, then the average response time $T(i)$ needs to be reduced by $(T\%(i) - T(i))$. Therefore, for $T\%(i)$ to approach $T\%bound$, the average response time target $Tt(i)$ becomes:

$$Tt(i) = T(i) - (T\%(i) - T\%bound(i)).$$

For the M/M/m queueing model,

$$Tt(i) \sim S(i) + p\{Nt(i)\}S(i) / (Nt(i) - R(i)S(i))$$

$$Tt(i) \sim S(i) + p\{N(i)\}S(i) / (N(i) - Rt(i)S(i))$$

and thus,

$$Nt(i) = \text{CEILING} [R(i)S(i) + p\{Nt(i)\}S(i) / (Tt(i) - S(i))]$$

$$Rt(i) = \text{FLOOR} [N(i)/S(i) - (p\{N(i)\}/Tt(i) - S(i))]$$

where $p\{m\}$ is the probability that there are m requests in the customer's server cluster

Table 4



For any given metric M ,

There are quick simulation tools, quick numerical computation tools and other approximation formula are available in computing $Nt(i)$ and $Rt(i)$ from given (i.e., measured) values of $R(i)$, $N(i)$ and $M(i)$.

Table 5